Adult Income Prediction Project Proposal

By

Razan Jaad

Data Science Bootcamp

SDAIA Academy

[September - December 2021]

1. **Question/need:**

This project Determines adult income based on a set of features. It analyzes the dataset and predicts whether the income of an adult will exceed 50k per year or not by developing a supervised machine learning model.

2. **Goal:**

In this project, we target the adult age.

3. **Data description:**

The dataset contains 48842 rows and 14 attributes with a size equal to 5,202KB. The target column is Income. This dataset is used to predict and determine whether a person makes over 50K a year, dataset contains information for an adult to determine the income if <=50k $ or not.

Dataset features:

1. Categorical Attributes: (contain only string values)

- workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
- education: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.
- marital-status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
- occupation: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.
- relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
- race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
- sex: Female, Male.
- native-country: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands.

2. Continuous Attributes: (contain only numerical values)

- age: continuous.
- fnlwgt: final weight, continuous.
  The weights on the CPS files are controlled to independent estimates of the civilian noninstitutional population of the US. These are prepared monthly for us by Population Division here at the Census Bureau.
- capital-gain: continuous.
- capital-loss: continuous.
- hours-per-week: continuous.
  Individual's working hours per week.

4. Algorithm:

The model used in this project is Logistic regression to predict whether a person makes over 50K a year.

5. Tools:

There are some libraries used in this project to analyze data, create models and data visualization:

- Python: High-level Software language.
- Jupyter notebook: Open-source web application.
- Sklearn: Software library
- Pandas: Software library
- Numpy
- Seaborn
- Matplotlib

And we will use machine learning and linear regression as well.

6. MVP goal:

Data preprocessing is a term used to describe any type of primary processing that applies to raw data. data. Over time, these techniques have evolved to involve users in data preparation. Train models for machine learning, artificial intelligence, and various data analysis. Can be used with various data sources (such as data stored in a database).[1]

Use the data from the system for preliminary treatment:[2]

- Handling Incomplete Data: That is, some cells contain empty values. Processing is:
    1) Ignore the entire data.
    2) Manually enter the value.
    3) Global constants are used to replace missing values with words such as unknown.
    4) Use the mean or median in the numeric field.

- Repeated data: Repeated values or datasets can be recognized by ready-made programs containing pre-packaged tools and code such as RapidMiner.

Then explore the dataset by drawing charts and displaying the results from the link below for the dataset:

https://archive.ics.uci.edu/ml/datasets/Adult

References:

- https://www.geeksforgeeks.org/data-preprocessing-in-data-mining/